

**SEX AND TISSUE EXPRESSION PATTERNS OF MEMBERS OF A FAMILY OF
GENE DUPLICATES IN *DROSOPHILA***

**A THESIS
PRESENTED TO THE FACULTY
OF CORNELL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE**

**BY
DORINA FRASHERI**

MAY 2013

© 2013 DORINA FRASHERI

ABSTRACT

This thesis focuses on the sex and tissue expression patterns of a family of reproductive proteins in *Drosophila*. This gene family was identified by Dr. Laura Sirot through a genomic screen of female reproductive proteins in *Drosophila melanogaster* that had undergone gene duplication. In most of the species analyzed, this family consists of three tandemly duplicated genes, which encode serine-type endopeptidase homologs. Through RT-PCR I show that two of the genes in this family are expressed in the female reproductive tract (RT), while the other is expressed in the male RT. Data across seven *Drosophila* species hint that this family likely arose from a single-copy gene that was initially female-specific; after duplication, one of the paralogs then evolved male-specific expression. A paralog's change in sex specificity (i.e. becoming co-opted for use by the opposite sex) is a unique case since in most instances of duplication in reproductive proteins the resulting paralogs continue to be expressed and function within the same sex. The reason co-option would take place here is unknown, but it provides for a means of molecular level communication between the sexes during mating given that the male-specific paralog is a seminal fluid protein.

BIOGRAPHICAL SKETCH

Dorina Frasheri was born in a small town in central Albania and is the third daughter of middle class parents, Shaban and Ronje Frasheri. Since a child Dorina had a keen interest in science and research, particularly math, astronomy, biology and medicine. Her childhood dream was to become an astronaut. However, living in Albania rendered the pursuit of scientific research almost impossible, so after finishing high school she decided to move to the United States in the hopes of a better education and a place that would provide her with broad opportunities for research.

Once in the United States, Dorina studied chemistry at St. John's University in Queens, New York, where she received her Bachelor of Science degree in 2006. She minored in philosophy and graduated with honors *summa cum laude*.

During college, she completed two summer internships at Bristol-Myers and Squibb Co., providing her with the first hands-on research opportunities and experiences.

Desiring more formal training, in August 2007 she pursued a Masters of Health Science degree in biochemistry and molecular biology at Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. She graduated in May 2008 and soon after worked for a few months in a clinical lab in Long Island, New York and later moved to the laboratory of Dr. Nicholas Restifo at the National Cancer Institute, focusing on the adoptive transfer of anti-tumor T cells in mice.

From August 2010 to August 2012 Dorina studied biochemistry, molecular and cell biology at Cornell University and here she presents her work performed in the laboratory of Dr. Mariana Wolfner.

To My Family

ACKNOWLEDGEMENTS

I would like to thank Dr. Mariana Wolfner for the chance to work in her lab and perform this research. I would also like to acknowledge Dr. Laura Sirot, who started this project, as well as the laboratory of Dr. Andrew Clark and the UCSD *Drosophila* Species Stock Center for providing the different *Drosophila* lines. I also would like to thank the Cornell Core facility for performing PCR sequencing as needed during this project. Last but not least, I would like to thank Dr. Thomas Fox and Dr. William Brown, who along with Dr. Mariana Wolfner and Dr. Andrew Clark, served as members of my committee.

TABLE OF CONTENTS

| | |
|--------------------------|------|
| BIOGRAPHICAL SKETCH..... | iii |
| DEDICATION..... | iv |
| ACKNOWLEDGEMENTS..... | v |
| LIST OF FIGURES..... | viii |
| LIST OF TABLES..... | ix |

| | |
|---|--------------------|
| <u>Chapter</u> | <u>Page</u> |
| I INTRODUCTION..... | 1 |
| II MATERIALS AND METHODS..... | 10 |
| 2.1 <i>Drosophila</i> Stocks..... | 10 |
| 2.2 Genomic Screen For Seminal Fluid Proteins (Sfps) Derived Via Duplication From Female Reproductive Proteins..... | 10 |
| 2.3 Sequence Retrieval And Identification Of Orthologs..... | 10 |
| 2.4 RNA Extraction..... | 11 |
| 2.5 RT-PCR..... | 14 |
| 2.6 Primers Used For PCR..... | 15 |
| 2.7 Genomic (gDNA) Preparations..... | 18 |
| III RESULTS..... | 20 |
| 3.1 Genomic Screen Of Twenty Secreted Female-Specific Reproductive Proteins In The Sperm Storage Organs Identifies Two Sets Of Such Proteins With Seminal Fluid Protein (Sfp) Paralogs..... | 20 |
| 3.2 Sex-Specific Or -Biased Expression Patterns Of Paralogs And | |

| | |
|--|----|
| Orthologs Across Seven <i>Drosophila</i> Species..... | 21 |
| 3.3 Tissue-Specific Or –Biased Expression Patterns Of Paralogs And | |
| Orthologs Across Three <i>Drosophila</i> Species..... | 28 |
| IV DISCUSSION AND CONCLUSION..... | 32 |
| V BIBLIOGRAPHY..... | 36 |

LIST OF FIGURES

| <u>Figure</u> | <u>Page</u> |
|---|-------------|
| Figure 1. Reproductive tracts of females and males from <i>D. melanogaster</i> , <i>D. ananassae</i> and <i>D. pseudoobscura</i> | 13 |
| Figure 2. Sex expression patterns of paralogs in <i>Drosophila melanogaster</i> | 22 |
| Figure 3. Sex expression patterns of paralogs in <i>Drosophila yakuba</i> | 23 |
| Figure 4. Sex expression patterns of paralogs in <i>Drosophila erecta</i> | 23 |
| Figure 5. Sex patterns of paralogs in <i>Drosophila sechellia</i> | 24 |
| Figure 6. Sex expression patterns of paralogs in <i>Drosophila simulans</i> | 24 |
| Figure 7. Sex expression patterns of paralogs in <i>Drosophila ananassae</i> | 25 |
| Figure 8. Sex expression pattern in <i>Drosophila pseudoobscura</i> | 25 |
| Figure 9. Chromosomal locations of all paralogs and orthologs..... | 26 |
| Figure 10. Phylogenetic tree of all paralogs and orthologs..... | 27 |
| Figure 11. Tissue expression patterns of paralogs in <i>Drosophila melanogaster</i> | 28 |
| Figure 12. Tissue expression patterns of paralogs in <i>Drosophila ananassae</i> | 29 |
| Figure 13. Tissue expression pattern in <i>Drosophila pseudoobscura</i> | 30 |

LIST OF TABLES

| <u>Table</u> | <u>Page</u> |
|--|-------------|
| Table 1. Genes in <i>Drosophila melanogaster</i> female sperm storage organs, plus the presence and identity of their seminal fluid protein paralogs..... | 20 |

CHAPTER I

INTRODUCTION

Gene duplication is thought to be the major driving force of biological novelty in evolution. Ohno's 1970 book *Evolution by Gene Duplication* puts forth the classical view that the evolution of genes and genomes is typically conservative in the absence of gene duplication, and that gene duplication plays a critical role in generating biodiversity (1). One of the earliest reported cases of gene duplication is in *Drosophila melanogaster*, where doubling of a chromosomal band in a mutant fly resulted in extreme reduction in eye size (the *Bar* mutation) (2). Ohno's book and, more recently, the emergence of genome sequencing and technology has led to an explosion of data on gene duplication. We now know that gene duplication is prevalent in all three kingdoms of life and that some species, such as *Saccharomyces cerevisiae* and *Arabidopsis thaliana*, have experienced several whole genome duplications (3-4), while most vertebrates have undergone only one or two large-scale genome duplications (5-7). These events have nonetheless shaped the genomes of the species by supplying raw genetic material.

While gene duplicates can become pseudogenes (unexpressed or non-functional genes), other outcomes such as neofunctionalization (the acquiring of a novel or related function by one duplicated copy) and subfunctionalization (where each gene resulting from duplication maintains part of the functions of the ancestral gene) can also occur (8-9). Pseudogenization, the process by which a duplicated gene becomes a pseudogene, removes a gene from the eyes of natural selection by silencing the gene through degenerative mutations, or by changing the gene's sequence sufficiently so that it is no longer identifiable as a duplicate of the original gene (10-12). On the other hand, neofunctionalization and subfunctionalization result in the retention of the duplicated genes. For a duplicated gene to undergo neofunctionalization, its new function

must be beneficial, so that the neofunctionalized copy will become preserved by natural selection (13). Subfunctionalization (where each duplicated copy becomes specialized to take on a subset of the functions of the ancestral gene, accumulating mutations that are beneficial for that function) represents itself in other forms: the most common consequence is a change in gene expression, where a gene is silenced in a tissue and/or developmental stage (14-18). Another form of it occurs when one of the duplicates is better than the parental gene at performing a subset of the latter gene's functions (19). Nonetheless, the evolutionary fate of a duplicated gene does not stop there. Gene duplication can also result in the conservation of gene function (i.e. neutralization). This can happen either through gene conversion, giving rise to genes that share very similar sequences and functions (i.e. concerted evolution) (20), or through purifying selection, which prevents duplicated genes from diverging by selecting against mutations that modify gene function (21-22). However, due to redundancy, two genes with identical functions are unlikely to be maintained in the genome for long, unless the duplication of gene product is advantageous to the organism.

What mechanisms give rise to duplication in the first place? Three mechanisms provide the opportunity for gene duplication: chromosomal or genome duplication, unequal crossing over, and retroposition. Chromosomal or genome duplication occurs when daughter chromosomes fail to disjoin after DNA replication. Unequal crossing over results in genes that are linked in a chromosome (i.e. tandem gene duplication). And duplication by retroposition occurs when an mRNA is retrotranscribed into cDNA and this cDNA is then inserted randomly into the genome (8-9, 23). Duplicated genes, whatever mechanisms and evolutionary forces they arise from, are referred to as paralogs and they form gene families. Paralogous genes are often clustered within a genome, although dispersed paralogs are also seen.

Duplication of genes provides an opportunity for co-option as well. Co-option is the process by which traits switch function, i.e. traits that had evolved under one set of conditions are co-opted to serve a different function under a second set of conditions. Co-option plays a crucial role in producing evolutionary innovations. Moreover, genetic co-option may involve more than just a change in function; it may involve a change in where a gene acts. For instance, α A crystallins, which are located in the lens of all vertebrates and which function to refract light passing through the lens, are derived from two subsequent gene duplication events. In the first duplication of a heat shock protein, a new heat shock protein and an α B crystallin (gene-duplicate) were formed. The new heat shock protein retained the plesiomorphic function (protection against damage) and plesiomorphic expression site (throughout the body). The α B crystallin protein, on the other hand, retained the expression pattern of its ancestral gene, but was co-opted for a new function (refraction of light). In the second duplication event (the duplication of the α B crystallin gene), a new gene encoding α B crystallin and one encoding α A crystallin were formed. The new α B crystallin retained the new function (refraction of light) and old location (expression throughout the body), while the α A crystallin maintained the new function but changed its expression pattern and was co-opted to solely function in the lens tissue (24).

Examples of gene duplication followed by co-option to serve in the reproductive arena include: serine proteases and their homologs, acid lipases, and odorant binding proteins, which are found in the seminal fluid of several fly species (25), but are thought to have originated from ancestral genes that initially were expressed throughout the body. These examples emphasize the ultimate story of gene duplication and co-option. Both phenomena provide the potential for rapid evolutionary change, and this in turn is evident in reproductive proteins.

Investigations of gene duplication in reproductive proteins have exploded in the last decade. The following are only a few of the many cases reported to date. One example is that of the reproductive protein Sp18 in abalone genus *Haliotis*. Sp18 is an 18kDa protein which is released from abalone sperm during the acrosome reaction and which mediates sperm-egg fusion. Sp18 appears to have arisen from a duplication event that also gave rise to lysin, a 16kDa abalone sperm protein that dissolves a hole in the vitelline envelope surrounding the egg, making it possible for the sperm to reach the egg plasma membrane. Lysin and Sp18 have such similar sizes, molecular weights, tertiary folds, and exon and intron arrangements, that even though their amino acid sequences are very different (sharing only 17-18% similarity, so that no BLAST search will match their extensively diverged sequences), they are thought to be descendants of a subfunctionalization duplication event (26-27). Another gene duplication event seems to have occurred to the sperm lysin of abalone *Haliotis tuberculata coccinea*. Both copies of the lysin protein resulting from duplication are expressed in the testis and both remain functional and have experienced positive selection. The paralogs share 88% identity in their coding sequences such that their proteins are 83% identical (with only 24 amino acid differences); the major amino acid sequence divergence between the two is in their C terminal regions (28).

Another instance of duplication in reproductive proteins is that of the cluster of three male-reproductive genes, *janusA*, *janusB*, and *ocnus* in *Drosophila melanogaster*, which lie within a 2.5-kb region of chromosome arm 3R. The genes are thought to have arisen from two tandem duplication events, the first of which happened around 35MY ago and duplicated *janusA*, via alternative splicing, into a new *janusA* and *janusB*. The new *janusA* gene had testis and general expression, whereas *janusB* had testis-specific expression. A subsequent duplication of *janusB* around 15MY ago created *ocnus*, whose expression is also testis-specific (29).

Other testis-specific reproductive proteins are the $\beta 2$ tubulins of the *Drosophila* and *Hirtodrosophila* species. These are highly conserved proteins (with no non-synonymous substitutions across 17 *Drosophila* and *Hirtodrosophila* species spanning 60 Myr of evolution) and function in generating the sperm-tail axoneme (30-31). They have evolved through gene duplication and subfunctionalization of the expression domain (32). In specific insects such as bees, wasps, etc., a second duplication event of the original conserved $\beta 2$ tubulin gene gave rise to new $\beta 2$ tubulin genes, each of which in turn has experienced more rapid evolution (33).

Another testis-specific gene, *k81*, arose from a duplication event of the *hiphop* gene in the *melanogaster* subgroup. The *hiphop* gene encodes a protein that binds to telomeres to prevent end-to-end fusion. K81 specifically marks telomeres in the male germ line and retains the telomere-protecting function of HipHop. But whereas HipHop functions in somatic cells, K81 is produced solely in males (34).

Interestingly, male-biased genes have arisen from duplication events more frequently than non-biased or female-biased genes. Moreover, a large number of the male-biased genes exhibit testis-specific expression (as evidenced in the above examples), and have evolved under positive selection. For instance, 83% of the nuclearly encoded mitochondrial genes in *Drosophila* show testis-specific expression, something that is not typical of the parental genes (35). Glycolytic enzymes in mammals exhibit testis-specific paralogs (36). A significant portion of the genes encoding the 26S *Drosophila* proteasome subunit show evolution of testis-specific expression (37). And last, but not least, two *Drosophila* nuclear transport proteins, Dntf-2 and Ran, have undergone duplication resulting in genes with testis-specific expression (38). How and why genes evolve to be expressed in such a specific tissue is not entirely known, but several mechanisms have been proposed: from DNA- and RNA-mediated relocation, to insertional biases, to testis-

specific regulatory elements, etc. Recently, a new mechanism was proposed by Gallach and Betran (39), saying that “intralocus sexual antagonism often begins in the parental gene via male selection for an allele (i.e. a sexually antagonistic allele) that performs better in testis, but worse in other tissues (i.e. male and female soma and ovaries) and culminates with the fixation of a relocated specialized male-specific gene.” They continue by explaining that because “testis is a tissue where sex-determination pathways are triggered and is very different from other tissues in that it must make sperm and is under strong selective pressures from sexual selection, parasite-related conflicts and segregation distortion to specialize and evolve quickly, testis probably generates most of the antagonistic conflicts among tissues, and consequently, it is likely the most sexually antagonistic tissue, thus explaining the amount of sex-biased genes expressed there and the amount of duplicated genes.” So, testis is not only the source of sexually antagonistic conflict, but also the place where this conflict gets resolved, leading to testis-specific expression. The testis may also release the genes from pleiotropic constraints and allow them to evolve more rapidly, consistent with the findings that male reproductive genes exhibit high rates of gene evolution (40).

Other cases of reproductive protein duplications include:

- (1) a chromosome 2L gene cluster of serine proteases in the female lower reproductive tract (RT) of the African malaria vector *Anopheles gambiae*, which are down-regulated by mating and may play a role in mating plug formation (41)
- (2) a three-paralog family encoding male-specific reproductive proteins (AgAcp34A1-3) in *Anopheles gambiae*, which play a role in sperm viability and function (42)
- (3) mammalian reproduction-related NLRPs (Nucleotide-binding oligomerization domain, Leucine rich Repeat and Pyrin domain containing proteins), some of which are oocyte-specific

and were duplicated before the divergence of mammals and then underwent a fast and independent functional diversification in different mammalian lineages providing mammals with reproductive advantages (43)

(4) in cows, the seminal ribonuclease gene, expressed in the tissues that make semen, is the result of a gene duplication event around 35MY ago that gave rise to the pancreatic ribonuclease gene as well (44-45)

(5) in rodents, seminal vesicle secretion (Svs) genes, which are important in the formation of the copulatory plug, have also arisen through duplication events (46-47)

(6) several digestive secreted proteases in *Drosophila arizonae* female lower RTs, particularly serine endoproteases, have resulted from recent, lineage-specific duplications. They have undergone rapid changes in their amino acid sequences and have experienced positive selection. This duplication and diversification possibly reflects the role of these enzymes in reproduction (likely the digestion of male seminal fluid proteins, Sfps) and the co-evolution of male and female proteins involved in reproduction (male proteins could respond to rapidly evolving female counterparts resulting in an intersexual arms race) (48-49)

(7) another protease gene family in *Drosophila mojavensis* (a five-paralog family), expressed in the female lower RT, has derived from recent gene duplications (50)

(8) a significant number of transferred Sfps in *Drosophila melanogaster* are encoded by genes that reside in clusters in chromosomes and are the result of tandem gene duplications (26)

(9) many male accessory gland proteins in *Drosophila mojavensis*/*Drosophila arizonae* have also arisen from gene duplications and show high rates of adaptive evolution (51)

(10) last but not least, the poly(A) polymerase GLD2 in *Drosophila melanogaster*, which is required for spermatogenesis, is a testis-specific expressed autosomal paralog of another poly(A)

polymerase, WISPY. WISPY is encoded by an X-linked gene and is expressed in the ovaries, where it is necessary for oogenesis and egg activation; in mice, nematode worms, and frogs, WISPY orthologs are also expressed in ovaries and (where tested) are required for oogenesis, suggesting that female germline expression/function is the ancestral state for this gene. *wispy* and *gld2* orthologs across several *Drosophila* species have maintained their expression patterns, adding to the importance of these genes in reproduction (52-54).

Analogous to the case of WISPY/GLD2, is the case presented in this thesis of three tandemly duplicated genes in *Drosophila melanogaster*, which encode serine-type endopeptidase homologs. These genes were found by Dr. Laura Sirot through a genomic screen aimed at identifying predicted-secreted, female-specific reproductive proteins with Sfp or male accessory gland protein paralogs. Twenty such female proteins, expressed in the female sperm storage organs, were screened. One pair (CG9897 and CG32834) was notable for its sequence similarity with CG32833, an Sfp. My data across seven *Drosophila* species suggest that this family of three genes, which reside in a tightly linked 4-kb cluster in chromosome 2R of *Drosophila melanogaster*, likely arose from a single-copy gene that was initially female-specific and that underwent two subsequent duplications, one resulting in a female-specific gene and the other in a male-specific gene. All genes are expressed in the RTs of the respective sex. Functional characterization of the three paralogs in *Drosophila melanogaster* shows that they play a role in female post-mating behaviors, such as re-mating receptivity, the number of eggs laid and the number of progeny produced (personal communication from Laura Sirot and Jessica Sitnik).

From all the evidence outlined above, it is clear that many male seminal fluid and female RT proteins, across several species, have been frequently subjected to gene duplication. A special case, such as the WISPY/GLD2 instance and the one presented here, is when a female RT

protein undergoes duplication and gives rise to a protein that changes its expression pattern, from female to male. This finding emphasizes the role that gene duplication plays in the acquisition of new genes (e.g. Sfps from female RT proteins) and diversification of closely related species.

CHAPTER II

MATERIALS AND METHODS

2.1 *Drosophila* Stocks

The Wolfner lab's Canton S strain was used for *Drosophila melanogaster*. *Drosophila yakuba* strain was purchased from UCSD Drosophila Species Stock Center (Cat #14021-0261.01). All other lines used were provided by the Clark laboratory at Cornell University. All lines were strains initially used for sequencing the genomes of respective species (12 Genomes Consortium 2007), and all were raised on standard yeast-glucose media at 22°C.

2.2 Genomic Screen For Seminal Fluid Proteins (Sfps) Derived Via Duplication From Female Reproductive Proteins

A genetic BLASTP screen was performed by Dr. Laura Sirot to search for *Drosophila melanogaster* female reproductive proteins that had paralogs in the male seminal fluid or accessory glands. Twenty proteins highly expressed in the female sperm storage organs were used for the screen (55-58). Genes resulting from the search were checked whether they represented an Sfp or male accessory gland protein, according to published data (26, 56, 59). Up to five hits per protein were selected. Hits were checked for paralogy by reciprocal BLAST comparisons. From this screen, two sets of female proteins expressed in the sperm storage organs that had paralogs in the seminal fluid or male accessory glands were identified. One set was chosen for further expression analysis.

2.3 Sequence Retrieval And Identification Of Orthologs

The *Drosophila melanogaster* genes were screened for orthologs across 11 *Drosophila* species, whose genome sequences were available (12 Genomes Consortium 2007). Via best reciprocal BLASTP, orthologs of one or all three genes were identified in the following species: *Drosophila simulans*, *Drosophila sechellia*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila ananassae*, *Drosophila pseudoobscura*, and *Drosophila persimilis*.

Drosophila simulans, *Drosophila sechellia*, *Drosophila yakuba*, and *Drosophila erecta* contain one ortholog of each gene from the set of three. *Drosophila ananassae* contains 4 copies (with GF11311 as the extra copy), while *Drosophila pseudoobscura* and *Drosophila persimilis* contain one copy only. No orthologs were identified in the more distant species. A *Drosophila persimilis* strain was not available for usage at the time, therefore it was not included in the analysis.

2.4 RNA Extraction

To test for sex-specific expression, total RNA was extracted from 10 female or 10 male flies of each species analyzed. The flies were homogenized with a pestle, in a microcentrifuge tube, in 150 uL TRIzol reagent (Invitrogen Cat #15596-026). Another 850 uL TRIzol, for a total of 1 mL, was added after homogenization. To remove fly parts and lipids content, the samples were centrifuged at 12,000 g for 10 minutes at 2-8°C. The supernatants were transferred to a new Eppendorf tube and incubated at room temperature for 5 minutes. 200 uL chloroform was added to each sample, which in turn was capped securely, shaken vigorously for 15 seconds, incubated at room temperature for 2-3 minutes, and centrifuged for 15 minutes at 12,000 g at 2-8°C. Following the spin, the top, clear, aqueous phase, which contained the RNA, and which was around 600 uL in volume, was carefully transferred (200 uL at a time) to a new Eppendorf tube. To precipitate the RNA, 500 uL isopropanol was added to each sample, and the sample was

then incubated at room temperature for 10 minutes, followed by centrifugation at 12,000 g for 10 minutes at 2-8°C. The supernatant was removed and the RNA, which formed a gel-like pellet on the bottom of the tube, was washed with 1 mL of 75% RNase-free ethanol, mixed by vortexing gently, and centrifuged at 7500 g for 5 minutes at 2-8°C. The supernatant was removed again and the RNA was air-dried for 5 minutes. Finally, the RNA was dissolved in a small volume of RNase-free water (~100 uL) by mixing up and down with a pipet, then incubated for 10 minutes at a 37°C water bath, and later stored at -20°C.

To test for tissue-specific expression, female and male flies from *Drosophila melanogaster*, *Drosophila ananassae* and *Drosophila pseudoobscura* were dissected into: the gonads, the rest of the reproductive tract, and the rest of the carcass (body without the reproductive tract (RT)). Ten flies of each species and each sex were used for RNA extraction from the carcass, whereas 50 flies of each species and each sex were used to extract RNA from the gonads and from the rest of the RTs. RNA extraction from the carcass was performed as outlined above. RNA extraction from the gonads and from the rest of the RTs were performed as follows:

The gonads and the rest of the RTs were dissected in 1x PBS (phosphate buffer saline). The dissected tissues were picked from the 1x PBS solution with tweezers and placed in an Eppendorf tube containing 100 uL TRIzol reagent. Tissues were homogenized with a pestle, and another 200 uL TRIzol was then added, bringing the total volume to 300 uL. The samples were then transferred into phase-lock tubes and incubated for 5 minutes at room temperature. This was followed by an addition of 100 uL chloroform, vortexing for 10 seconds, and another 5-minute incubation at room temperature. The samples were then centrifuged at 13,200 rpm for 15 minutes at 4°C. The supernatants, which contained the RNA, were placed in new Eppendorf tubes. To precipitate the RNA, 75 uL isopropanol was added to each sample. Samples were then shaken by

hand for 10 seconds and let sit at room temperature for 10 minutes. The samples were stored overnight at -20°C. The next day, samples were thawed for a few minutes and then centrifuged at 11,000 rpm for 15 minutes at 4°C. The supernatants were discarded and the pelleted RNA was washed with 250 uL of 75% RNase-free ethanol. The samples were again centrifuged as in the previous step, the supernatants were again removed, and the samples were air-dried for 5-10 minutes. Finally, the RNA was dissolved in a small volume of RNase-free water (~10 uL) by mixing up and down with a pipet, and later stored at -20°C.

Below, in Figure 1, are pictures of the RTs of the 3 species dissected for the tissue expression experiments.

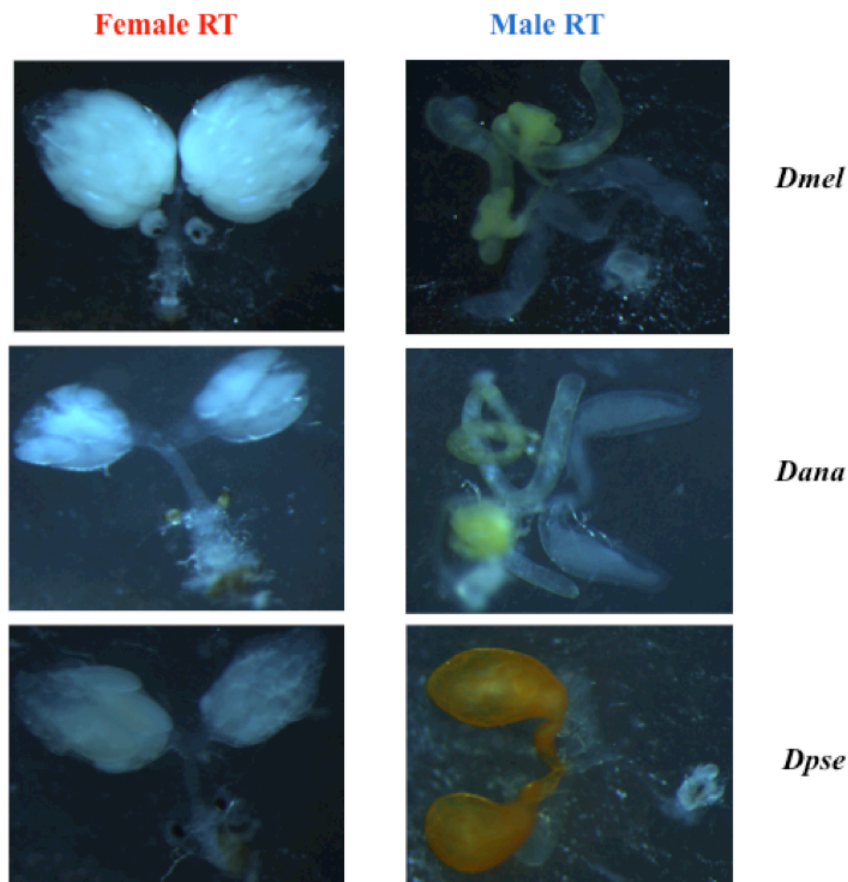


Figure 1. Pictures of the reproductive tracts (RTs) of females and males from *Drosophila melanogaster*, *Drosophila ananassae* and *Drosophila pseudoobscura*. Female gonads are easily distinguished due to

their bulky white shapes. Male gonads are yellow in *Drosophila melanogaster* and *Drosophila ananassae*, and orange in *Drosophila pseudoobscura*. *Drosophila ananassae* RTs of both sexes are much smaller than *Drosophila melanogaster* ones, as is the ejaculatory duct of male *Drosophila pseudoobscura*.

2.5 RT-PCR

To remove any residual genomic DNA prior to RT-PCR, 1 ug of total extracted RNA was treated with 1 uL RQ1 DNase enzyme (RNase-free) in the presence of RNase-free water and 10x RNase-free DNase Buffer for a total of 10 uL reaction, and incubated at 37°C for 30 minutes. To terminate the reaction, 1 uL DNase Stop Solution was added and the mixture was incubated for 10 minutes at 65°C. All reagents were obtained from Promega, Part# 9PIM610.

A portion of the DNase-treated RNA (4 uL or 0.4 ug) was used to synthesize first-strand cDNA with 1 uL oligoDT primer stock (at 20uM). The mixture (5 uL total) was then heated at 72°C for 3 minutes, and later placed immediately on ice. To this mixture were then added the following reagents: 2 uL 5x First-Strand Buffer, 1 uL 10mM dNTP mixture, 1 uL 20mM DTT, and 1 uL SMART-Scribe RTase, for a total reaction of 10 uL. The reaction was incubated at 42°C for 70 minutes. To terminate the reaction, the samples were heated at 70°C for 15 minutes. Finally the samples were stored at 2-8°C so they could later be used for gel electrophoresis. All reagents were obtained from Clontech, #PT4080-2.

The resulting cDNA was diluted 10-fold with RNase-free water and 1-3 uL of it were used in subsequent PCR reactions to test for gene expression. Standard PCR reactions consisted of: 5 uL 5x GoTaq Green Buffer, 1 uL 10mM dNTPs, 1-3 uL diluted cDNA, 1 uL of each forward and reverse primer (at 10uM concentration), 0.25 uL GoTaq Enzyme, and RNase-free water, for a

total of 25 uL. PCR products and expression patterns were visible with 30-35 cycles of amplification.

For visualization, samples were run on 1% agarose gel electrophoresis and were stained with 3x GelRed (diluted from the 10,000x stock solution using water). GelRed 10,000x stock solution was obtained from Biotium Products Cat #41003.

2.6 Primers Used For PCR

[Flybase.org](http://flybase.org) was used to retrieve the coding sequences of all orthologs. Then PCR primers were designed using the Primer3 program v0.4.0 (<http://frodo.wi.mit.edu>). Primers were designed such that ~350 bp regions would be produced. The selected primers were then checked against the UCSC genome database to make sure that they would not amplify any other region besides the one desired. RpL32 (ribosomal protein L32) was used as a positive control. RpL32 primers span an intron-containing region, allowing the chance to distinguish between the genomic DNA (larger band) and the cDNA (smaller band) and to insure that no contamination occurred between the two. The following are the primers used for each ortholog in each species throughout the study:

DanaGF11310 F: TGTCGCCTCTGTACATCCTG T_m = 56.5°C
DanaGF11310 R: TTGAGCACTTTTGCTTGCTG T_m = 54.4°C
Length CDS: 345 bp
Length genomic: 345 bp

DanaGF11311 F: GGCACGTCAGTTCGTAACAC T_m = 56.3°C
DanaGF11311 R: GCTTCTTCCTCTTGGTGGTG T_m = 55.8°C
Length CDS: 352 bp
Length genomic: 352 bp

DanaGF11312 F: AGGTGGAGGCCATAAAGGTC T_m = 56.8°C
DanaGF11312 R: GCTTCCAGACCTCCGTACAC T_m = 57.5°C
Length CDS: 348 bp
Length genomic: 348 bp

DanaGF11314 F: AGTGCCGGACAACGATAAAC T_m = 55.4°C
 DanaGF11314 R: CAGCTCTTGAATCCCGCTAC T_m = 55.6°C
 Length CDS: 348 bp
 Length genomic: 348 bp

DereGG20079 F: GCGCCATTATCTCGAAGAAC T_m = 53.8°C
 DereGG20079 R: TTCCGCTGATAAGCCTCTTG T_m = 54.8°C
 Length CDS: 345 bp
 Length genomic: 345 bp

DereGG20080 F: AAACGATCCAGACCATCCAG T_m = 54.5°C
 DereGG20080 R: GTAGACTTCAGGCCGTTTGG T_m = 55.7°C
 Length CDS: 350 bp
 Length genomic: 350 bp

DereGG20082 F: CGGATGATGATACCATGTGC T_m = 53.4°C
 DereGG20082 R: CTATTCTCGGCAATCCAACC T_m = 53.4°C
 Length CDS: 355 bp
 Length genomic: 355 bp

DpseGA25104 F: AGCTTGTGCCTCTTGTGGAG T_m = 57.6°C
 DpseGA25104 R: CTGCGCCAGATTAGCATAGAC T_m = 55.6°C
 Length CDS: 345 bp
 Length genomic: 345 bp

DsecGM15594 F: GAGATCAGCGGATCATAAACG T_m = 53.2°C
 DsecGM15594 R: GTGCTTTGTTGGCCCTCTC T_m = 56.7°C
 Length CDS: 345 bp
 Length genomic: 345 bp

DsecGM15595 F: ATTCAGAGGAGGACGACAGC T_m = 56.5°C
 DsecGM15595 R: TTGCCTGGATCTTTTGGAG T_m = 53.0°C
 Length CDS: 351 bp
 Length genomic: 351 bp

DsecGM15596 F: GAACTGGATTCCTGCTCGAC T_m = 55.5°C
 DsecGM15596 R: GAGGGTGAGGTACGAGATGC T_m = 57.1°C
 Length CDS: 347 bp
 Length genomic: 347 bp

DsimGD15206 F: GCTTCGTGTGTCCAGTCCTAC T_m = 57.6°C
 DsimGD15206 R: ACTGCTCCCGATTGTAGACG T_m = 56.7°C
 Length CDS: 352 bp
 Length genomic: 352 bp

DyakGE11617 F: ATTGAAAACCTGCGAGCAAC T_m = 54.2°C
DyakGE11617 R: CTCCTTGTCTGAGGAGCAG T_m = 57.1°C
Length CDS: 347 bp
Length genomic: 347 bp

DyakGE11618 F: AAGAGGACGACGACTGCAAC T_m = 57.3°C
DyakGE11618 R: GTTTGGCCAACTGGATTGTC T_m = 54.6°C
Length CDS: 354 bp
Length genomic: 354 bp

DyakGE11619 F: TTGGAGGCTATGATGTGGAC T_m = 54.3°C
DyakGE11619 R: CATAGCACCAGGAATCATCG T_m = 53.1°C
Length CDS: 354 bp
Length genomic: 354 bp

DmelCG32834 F: ATTCAGCCCATCAGCATAGC T_m = 55.3°C
DmelCG32834 R: TGGCGTAGACATCTGGTTTG T_m = 54.9°C
Length CDS: 349 bp
Length genomic: 349 bp

DmelCG32833 F: ACGACTCTTGGTGGTCATCC T_m = 56.8°C
DmelCG32833 R: ATTAGACGGAAGCTGGTTGG T_m = 54.8°C
Length CDS: 345 bp
Length genomic: 345 bp

DmelCG9897 F: TGCGAGCTACTCAACACCAC T_m = 57.4°C
DmelCG9897 R: AGACATCGGGCTTAATGGAG T_m = 54.3°C
Length CDS: 349 bp
Length genomic: 404 bp

RpL32-5spp-F: CGCACCAAGCACTTCATC T_m = 54.4°C
RpL32-5spp-R: GGTGCGCTTGTTTCGATCC T_m = 56.9°C
Length CDS: 153 bp
Length genomic: 215 bp

Due to a well-conserved sequence, this last set of primers (for the positive control RpL32 gene) worked for the following species: *Drosophila melanogaster*, *Drosophila simulans*, *Drosophila sechellia*, *Drosophila yakuba*, and *Drosophila erecta*. However, specific primer sets were needed, and designed, for *Drosophila pseudoobscura* and *Drosophila ananassae* RpL32 genes.

RpL32-pse-F: TCACCAGTCGGATCGTTATG T_m = 54.2°C

RpL32-pse-R: TATGACGGGTACGCTTGTTG $T_m = 54.9^{\circ}\text{C}$
Length CDS: 139 bp
Length genomic: 212 bp

RpL32-ana-F: GCCCAAGATCGTTAAGAAGC $T_m = 53.8^{\circ}\text{C}$
RpL32-ana-R: TTGGGCATCAGGTACTGACC $T_m = 56.9^{\circ}\text{C}$
Length CDS: 144 bp
Length genomic: 208 bp

In order to identify the sex expression of *Drosophila simulans* unannotated 9897 and 32833 gene paralogs, *Drosophila sechellia* GD15594 and GD15595 primer sets were used respectively. *Drosophila sechellia* is closely related to *Drosophila simulans*, thus the same primer sets were anticipated to work very well for both.

2.7 Genomic DNA (gDNA) Preparations

To prepare gDNA, 4 flies (mix of females and males) from each species were placed in an Eppendorf tube and kept on ice. To break down the cell membranes and free the DNA, 100 μL of a solution containing 0.1M Tris HCl/0.1M EDTA pH 9.0, 1% SDS and 1% DEPC in ethanol, was added to each tube. Samples were homogenized with a pestle until no recognizable fly parts were visible, and then incubated at 70°C for 30 minutes. To precipitate proteins and the SDS, 14 μL of 8M KAc (potassium acetate) was added to each homogenate, followed by a 30-minute incubation on ice. Samples were then centrifuged at 14,000 rpm for 15 minutes at 4°C . The supernatants, which contained the DNA, were transferred to new Eppendorf tubes. To precipitate the DNA, isopropanol (at 0.5 the volume of the collected sample) was added to each tube. The samples were incubated at room temperature for 15 minutes, and then centrifuged at 14,000 rpm for 5 minutes at room temperature. The supernatants were discarded and the pellets were washed with 200 μL of 70% RNase-free ethanol. The samples were later briefly vortexed and centrifuged

for 1-2 minutes at room temperature at 14,000 rpm. The supernatants were again discarded and the samples were air-dried for 5-10 minutes. Finally, the pellets were dissolved in a small volume of RNase-free water (~30 uL) by mixing up and down with a pipet, and later stored at -20°C. gDNA was diluted 10-fold with RNase-free water when ready for PCR.

CHAPTER III

RESULTS

3.1 Genomic Screen Of Twenty Secreted Female-Specific Reproductive Proteins In The Sperm Storage Organs Identifies Two Sets Of Such Proteins With Seminal Fluid Protein (Sfp) Paralogs

A targeted search for secreted female-specific reproductive proteins with duplicates that were male-expressed Sfps found 2 such sets of proteins (personal communication from Laura Sirot). The 20 female-specific proteins used in the search are expressed in the sperm storage organs (Table 1 below). The first set with sex-switched duplicates, included lipases, YP1, YP2, and YP3, which are expressed in the spermathecae (the female sperm storage organs) and fat body, and show sequence similarity to CG5161, an Sfp. The second set consisted of CG9897 and CG32834, which are expressed in the spermathecae and show evidence of homology with the Sfp, CG32833, which is expressed in the male accessory glands (according to microarray data for the three genes (56)).

The second set was chosen for further expression and functional analysis, and evolutionary history investigation.

| Family | Gene | Class | SFP paralog |
|---------------------------------------|--|-----------------|------------------------|
| Spermathecal endopeptidases (SEND) | CG17012 (SEND1) CG17234 CG17239 CG17240 (ser12) CG18125 (SEND2) CG31861 | Serine protease | None |

| | | | |
|--|---|--|---------|
| Inactive spermathecal endopeptidases (ISEND) | CG9897 CG32834 | Inactive serine protease | CG32833 |
| Yolk Protein | CG2985 (YP1) CG2979 (YP2) CG11129 (YP3) | Lipase | CG5162 |
| Other | CG6426 | Destabilase | None |
| | CG13318 | Serine protease | None |
| | CG18067 | 3',5'-cyclic-nucleotide phosphodiesterase activity | None |
| | CG18525 | Serine protease inhibitor | None |
| | CG18628 | No conserved domains | None |
| | CG30371 | Serine protease | None |
| | CG31686 | No conserved domains | None |
| | CG32277 | Serine protease | None |
| | CG32751 | Hydrolase | None |

Table 1. Genes highly-expressed in *Drosophila melanogaster* female sperm storage organs, plus the presence and identity of the seminal fluid protein paralogs (taken from Sirot *et al.*, in prep).

3.2 Sex-Specific Or -Biased Expression Patterns Of Paralogs And Orthologs Across Seven *Drosophila* Species

To confirm the sex expression patterns of the *Drosophila melanogaster* paralogs, as published previously (56), I performed RT-PCR from whole animal female and male flies (as described in

MATERIALS AND METHODS). I confirmed that CG9897 shows female-biased expression, whereas CG32833 and CG32834 show male- and female-specific expression, respectively.

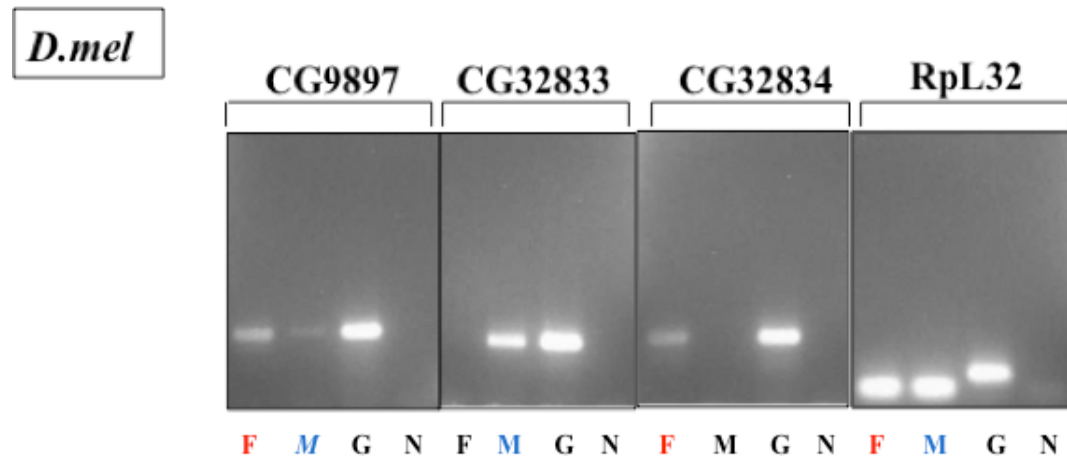


Figure 2. Sex-specific or -biased expression patterns of paralogs in *Drosophila melanogaster*. F means female. M means male. Labels in italics mean slight expression in that particular sex. G is genomic DNA. N is the negative control (H₂O). RpL32, ribosomal protein L32, is the positive control.

The number of paralogs (three), their order and sex-of-expression patterns were well conserved from *Drosophila melanogaster* to *Drosophila yakuba*, *Drosophila erecta*, *Drosophila sechellia*, and *Drosophila simulans* (i.e. the *melanogaster* subgroup) as shown below in Figures 3-6. The few exceptions are: *Drosophila erecta* GG20079 (the ortholog of CG9897) shows female-specific expression, not female-biased expression, whereas GG20082 (the ortholog of CG32834) shows female-biased expression, not female-specific expression; also, GM15594 in *Drosophila sechellia* (the ortholog of CG9897) shows female-specific expression, not female-biased expression; and *Drosophila simulans* GM15595 (the ortholog of CG32833) shows male-biased expression, not male-specific expression.

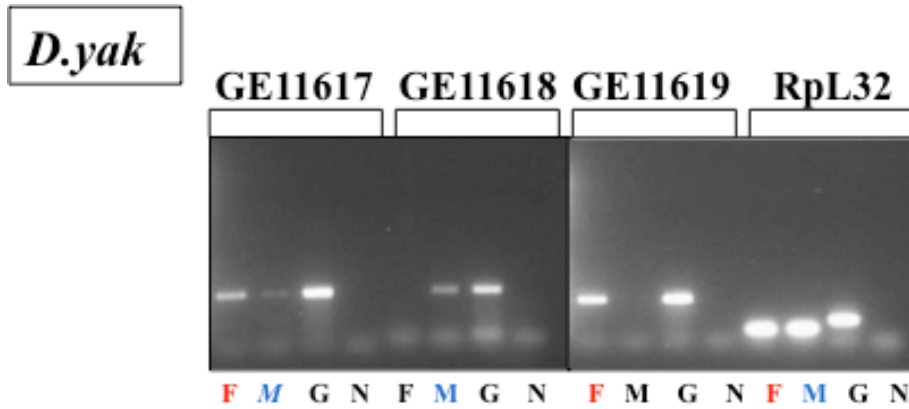


Figure 3. Sex-specific or –biased expression patterns of paralogs in *Drosophila yakuba*. F means female. M means male. Labels in italics mean slight expression in that particular sex. G is genomic DNA. N is the negative control (H₂O). RpL32, ribosomal protein L32, is the positive control.

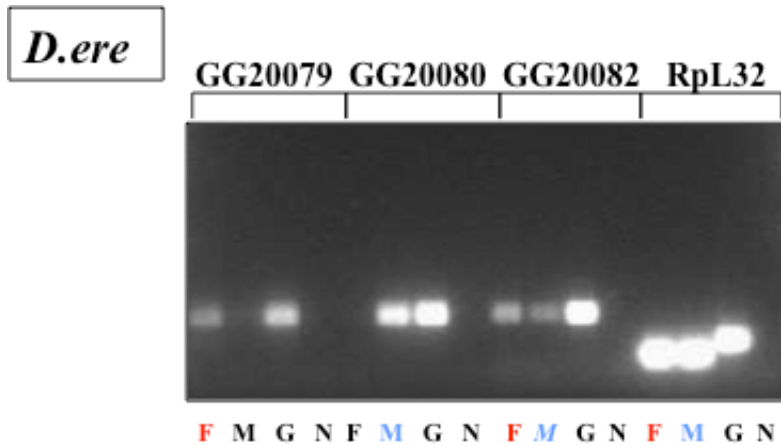


Figure 4. Sex-specific or –biased expression patterns of paralogs in *Drosophila erecta*. F means female. M means male. Labels in italics mean slight expression in that particular sex. G is genomic DNA. N is the negative control (H₂O). RpL32, ribosomal protein L32, is the positive control.

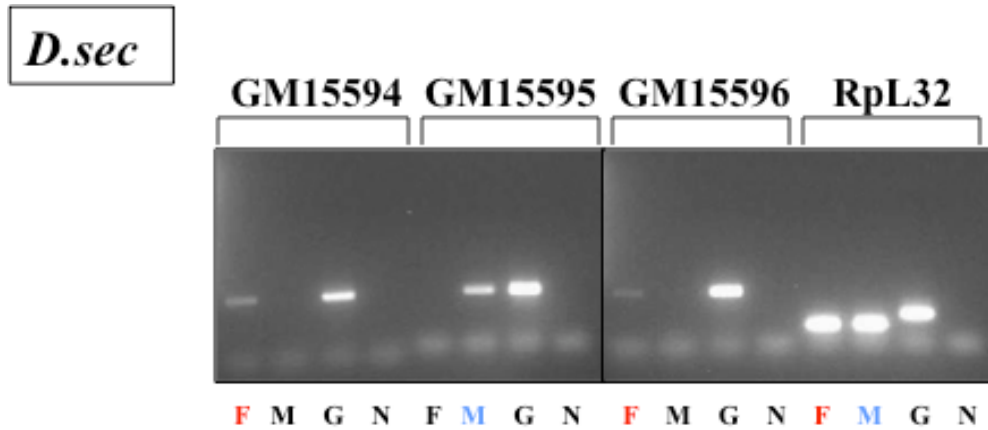


Figure 5. Sex-specific expression patterns of paralogs in *Drosophila sechellia*. F means female. M means male. G is genomic DNA. N is the negative control (H₂O). RpL32, ribosomal protein L32, is the positive control.

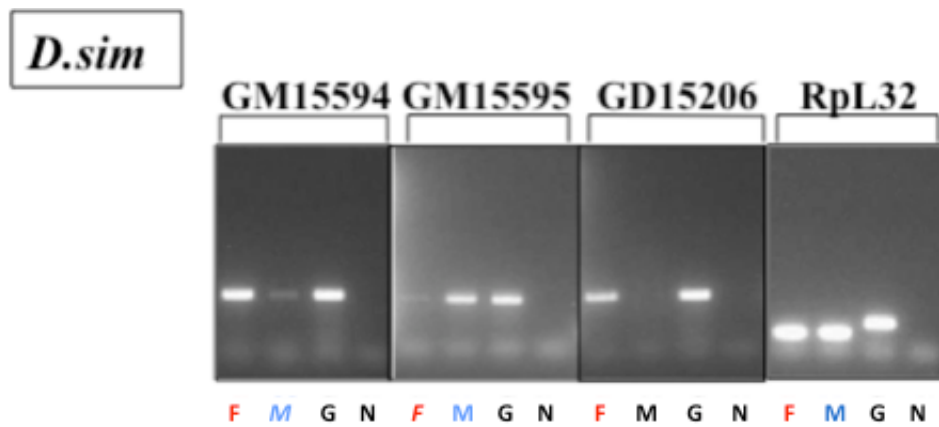


Figure 6. Sex-specific or -biased expression patterns of paralogs in *Drosophila simulans*. F means female. M means male. Labels in italics mean slight expression in that particular sex. G is genomic DNA. N is the negative control (H₂O). RpL32, ribosomal protein L32, is the positive control.

Drosophila ananassae has 4 copies of these genes, meaning that the 4th copy possibly arose from another round of gene duplication. The following are their sex expression patterns.

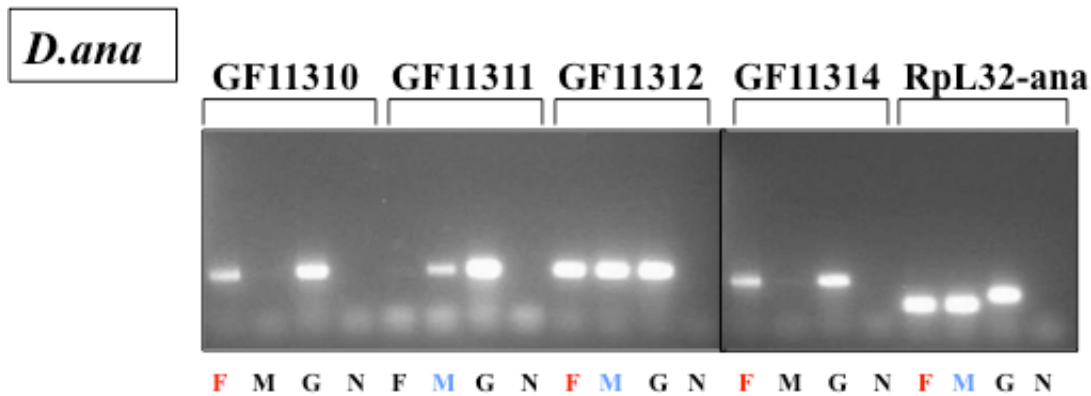


Figure 7. Sex-specific or -biased expression patterns of paralogs in *Drosophila ananassae*. F means female. M means male. Labels in italics mean slight expression in that particular sex. G is genomic DNA. N is the negative control (H₂O). RpL32-ana, ribosomal protein L32 in *ananassae*, is the positive control.

The only existing paralog in *Drosophila pseudoobscura* shows female-specific expression (Figure 8 below), suggesting that before duplication, the original gene was female-specific in expression. Following duplication, one copy maintained the female expression, while the other (possibly arising from a second duplication event) was co-opted to be expressed and function in the males.

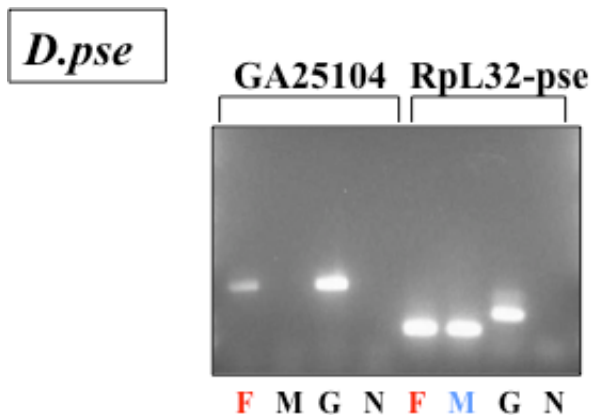


Figure 8. Sex-specific expression pattern of the single paralog in *Drosophila pseudoobscura*. F means female. M means male. G is genomic DNA. N is the negative control (H₂O). RpL32-pse, ribosomal protein L32 in *pseudoobscura*, is the positive control.

A summary of the chromosomal locations of all the paralogs and orthologs as well as their expression patterns in a phylogenetic tree are shown below in Figures 9-10.

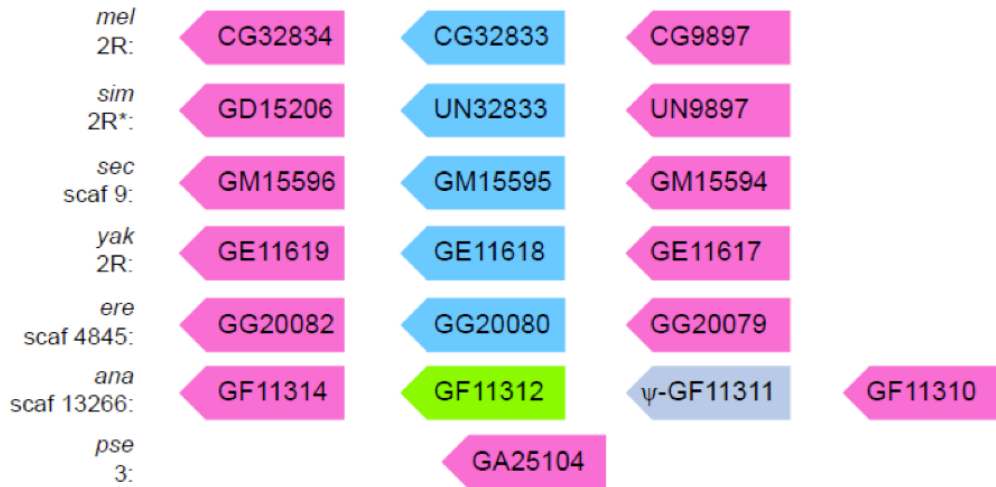


Figure 9. Chromosomal locations of paralogs and orthologs across seven *Drosophila* species as outlined in flybase.org. Pink represents female-specific or –biased expression, blue represents male-specific expression, light blue means very low expression in males, green means unbiased expression in both sexes (at approximately equal levels), and Ψ means likely a pseudogene. The first three letters indicate the *Drosophila* species (*mel*: *melanogaster*; *sim*: *simulans*; *sec*: *sechellia*; *yak*: *yakuba*; *ere*: *erecta*; *ana*: *ananassae*; *pse*: *pseudoobscura*). These letters are followed by the chromosomal locations of the gene clusters in each species. In *Drosophila simulans*, GD15206 is found in an unassembled part of chromosome 2R (indicated by the asterisk), while UN32833 and UN9897 represent unannotated copies whose sequences were determined by sequencing or BLAST. The Dsim\UN32833 sequence is only partially determined (the 142 codons at the start of the coding sequence) (taken from Sirot *et al.*, in prep).

Sequencing of *Drosophila ananassae* GF11311 showed that this gene is mis-annotated: it does not contain a 21-bp intron as predicted, and there is no evidence of splicing when gDNA and cDNA amplified PCR products are compared. The gene actually contains a premature stop

codon, thus it is likely to be non-functional (a pseudogene) (Sirot *et al.*, in prep). Moreover, when compared to all other genes in this family, GF11311 required 3x more template cDNA and 5 more cycles for its PCR reaction to have robust amplification (as shown in Figure 7), suggesting that the expression of this gene in males is at a very low level.

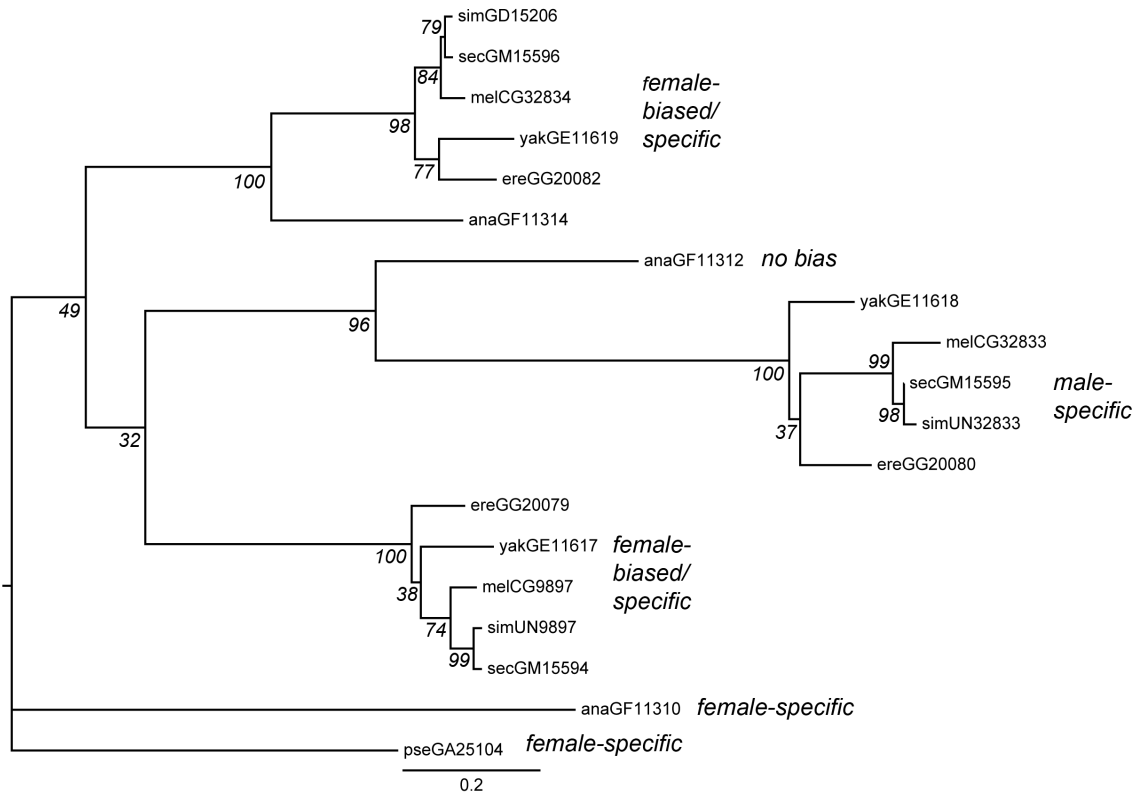


Figure 10. Phylogenetic tree of protein sequences of the paralogs and orthologs across seven *Drosophila* species. Tip labels indicate protein names; the first three letters indicate the *Drosophila* species (mel: *melanogaster*; sim: *simulans*; sec: *sechellia*; yak: *yakuba*; ere: *erecta*; ana: *ananassae*; pse: *pseudoobscura*), and the following characters indicate the FlyBase gene name. “UN” in the gene name indicates a previously unannotated copy of the gene in *Drosophila simulans*. Scale bar indicates the number of substitutions per site. Calls of orthology are consistent with phylogenetic clustering and gene order (see Figure 8): the six genes shown at the top of the figure (GD15206-GF11314) are one set of orthologs, GF11312-GG20080 are another set, and GG20079-GF11310 are the third set. The tree is

rooted on the single *Drosophila pseudoobscura* copy of this gene family, GA25104. Expression patterns are indicated in *italics* text (taken from Sirot *et al.*, in prep).

As evidenced from the tree, male-specific genes have undergone an accelerated rate of amino acid substitution. Finally, GF11311, though not shown in the tree because it is likely a pseudogene, falls in the same middle branch as GF11312 and the male-specific genes.

3.3 Tissue-Specific Or –Biased Expression Patterns Of Paralogs And Orthologs Across Three *Drosophila* Species

After confirming the overall sex-expression patterns of the *Drosophila melanogaster* genes, and identifying those patterns in the other six *Drosophila* species, it was important to identify where in the body each gene was expressed. Published microarray data on *Drosophila melanogaster* places CG9897 in the female reproductive tract (RT), CG32933 in the male RT, and CG32834 in the female RT with a slight expression in the male RT (56). Through RT-PCR from dissected female and male flies (as described in MATERIALS AND METHODS), I confirmed the published data (as shown in Figure 11).

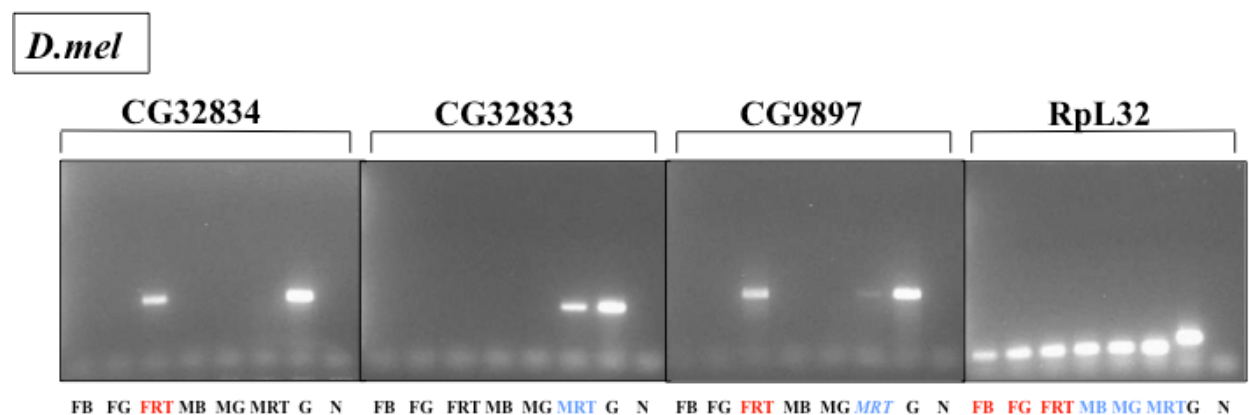


Figure 11. Tissue-specific or –biased expression patterns of paralogs in *Drosophila melanogaster*. FB or MB means female or male carcass (body without the reproductive tract). FG or MG means female or male

gonads. FRT or MRT means female or male reproductive tracts (without the gonads). Labels in italics mean slight expression in that particular tissue. G is genomic DNA. N is the negative control (H₂O). RpL32, ribosomal protein L32, is the positive control.

Furthermore, I was interested in checking the tissue expression patterns in two outlier species: *Drosophila ananassae* (since it contains one extra duplicated copy) and *Drosophila pseudoobscura* (the most distant species, and most likely the species where the first duplication event happened). Because dissecting the RTs and gonads from the rest of the bodies is labor intensive, only these 3 species were checked for tissue expression patterns. In *Drosophila ananassae* all 4 genes are expressed in the RTs; in females (GF11310, GF11314), males (GF11311), or both (GF11312) (Figure 12 below). In *Drosophila pseudoobscura*, the only existing paralog is expressed in the female RT (Figure 13 below).

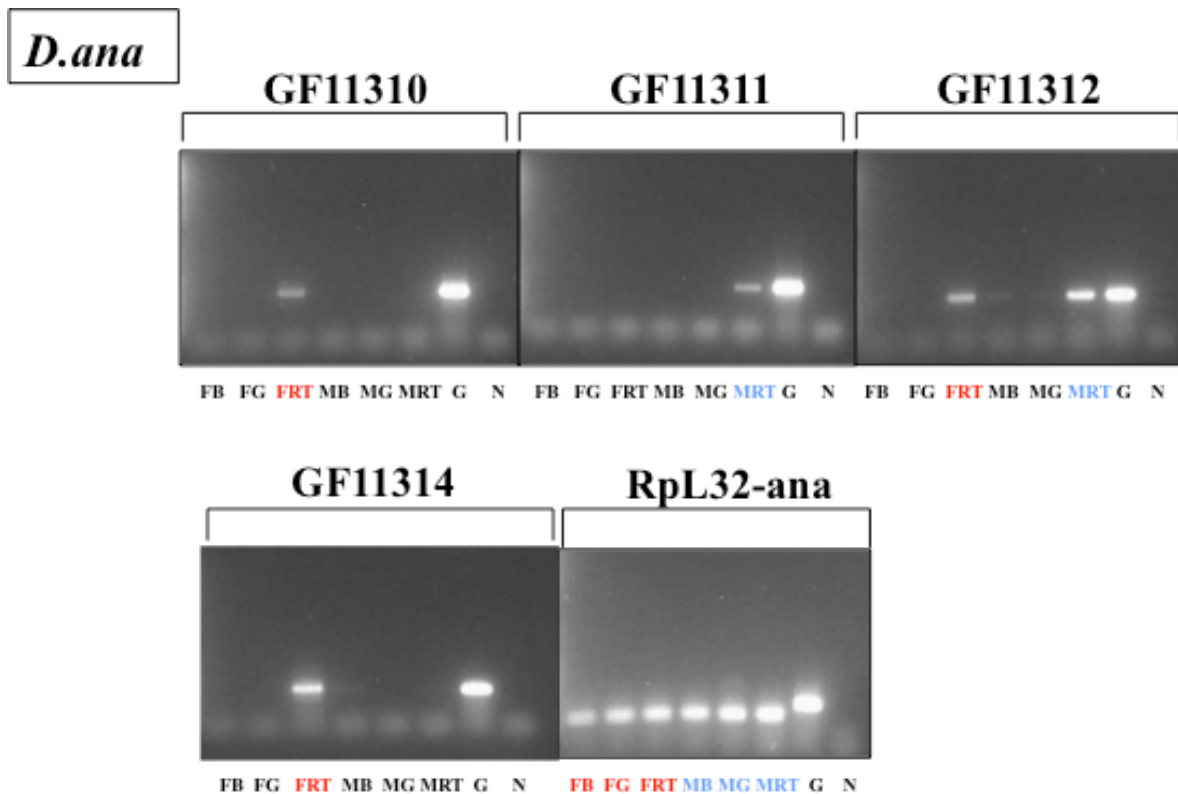


Figure 12. Tissue-specific or –biased expression patterns of paralogs in *Drosophila ananassae*. FB or MB means female or male carcass (body without the reproductive tract). FG or MG means female or male gonads. FRT or MRT means female or male reproductive tracts (without the gonads). Labels in *italics* mean slight expression in that particular tissue. G is genomic DNA. N is the negative control (H₂O). RpL32-ana, ribosomal protein L32 in *ananassae*, is the positive control. The light bands visible in the MB and MG of GF11312 and MB of GF11314 could be due to tissue carry over during dissection.

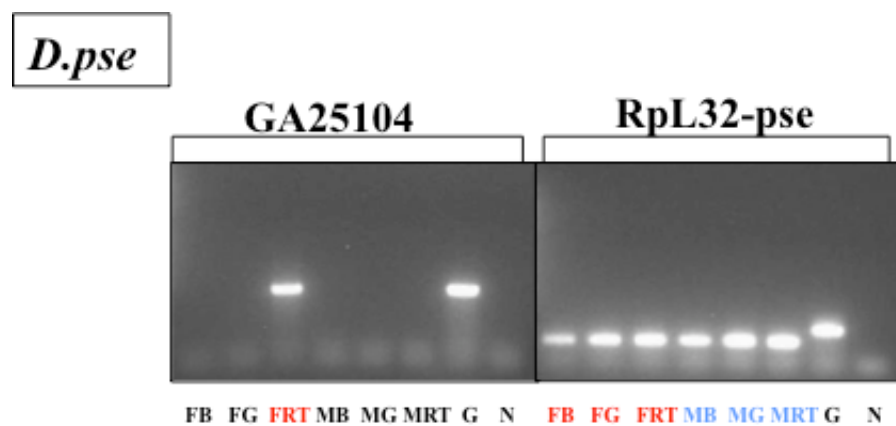


Figure 13. Tissue-specific expression pattern of the single paralog in *Drosophila pseudoobscura*. FB or MB means female or male carcass (body without the reproductive tract). FG or MG means female or male gonads. FRT or MRT means female or male reproductive tracts (without the gonads). G is genomic DNA. N is the negative control (H₂O). RpL32-pse, ribosomal protein L32 in *pseudoobscura*, is the positive control.

The evidence shows that the expression of these genes is limited to the reproductive tracts of female and male flies, suggesting that while the sex expression has changed following duplication (from female to male), the functional signature has remained the same (the genes are involved in reproductive aspects, at least in *melanogaster* where the functional analysis has been performed so far (personal communication from Laura Sirot and Jessica Sitnik). It is likely that

these genes perform similar reproductive functions in the other *Drosophila* species, considering that their expression patterns are very well conserved.

CHAPTER IV

DISCUSSION AND CONCLUSION

Gene duplication is an important mechanism for generating evolutionary novelty. Duplication of genes encoding reproductive proteins is particularly important because it can expand the suite of such proteins that are expressed within a sex. Recently, the case of WISPY/GLD2 proteins highlighted another facet of duplication of reproductive proteins, i.e. it can give rise to proteins that are expressed in the opposite sex. WISPY is expressed in the female germline and is necessary for oogenesis and egg activation, while GLD2, the gene that arose via a duplication event of WISPY, is expressed in the male germline and is necessary for male fertility and regulates mRNAs in spermatogenesis (52-54). Analogous to WISPY/GLD2 is the case of the gene family presented in this thesis. In most of the species analyzed, this family consisted of three tandemly duplicated genes, which encode serine-type endopeptidase homologs. Through RT-PCR in three *Drosophila* species I showed that two of the genes in this family are expressed in the female reproductive tract (RT), while the other is expressed in the male RT. Whole animal expression-pattern data across seven *Drosophila* species suggested that this family likely arose from a single-copy gene that was initially female-specific (due to the fact that the single copy in *Drosophila pseudoobscura* is female-specific); after the divergence of *Drosophila pseudoobscura* and the *melanogaster* group species, two duplication events occurred: one giving rise to a male-specific paralog, the other giving rise to a paralog that maintained the female-specific expression pattern. A possible additional lineage-specific duplication event occurred in *Drosophila ananassae*, giving rise to GF11311. However, it is likely that this gene is a pseudogene (Sirot *et al.*, in prep).

The case of this gene family is even more interesting than WISPY/GLD2 since the male paralog is a seminal fluid protein (Sfp), which is transferred to females upon mating and affects post-mating behavior. Indeed, functional analysis on the three genes in *Drosophila melanogaster* suggests that each is required for normal post-mating responses in females (personal communication from Laura Sirot and Jessica Sitnik). Three post-mating behaviors were tested in the *melanogaster* species: the probability of remating, the number of eggs laid, and the number of adult progeny. The probability of remating was tested as follows: control or knockdown females were mated once to control or knockdown males, then tested whether they would remate with a wild-type (Canton S strain) male within a 1-hour time period at 1, 4, or 10 days after the initial mating. The number of eggs laid and the number of adult progeny were measured in the first 24-hours after mating and over a 10-day period after mating. Female genes were knocked down in the spermathecae with the spermathecal-specific *Send1*-GAL4 driver. CG32833 was knocked down in the male accessory glands with the male accessory gland specific *ovulin*-GAL4 driver. Tissue-specific GAL4 drivers were used because the genes show tissue-specific expression and because the *tubulin*-GAL4 driver caused lethality when used to knock down CG32833.

The findings were that knockdown of any of the members of this gene family caused females to be less receptive to remating. This effect was greatest at the latest time-point, 10 days after the initial mating. Also, CG32834-knockdown females laid fewer eggs over the course of 10 days and especially in the first 24 hours after mating than control females. However, egg-laying was not affected by knocking down CG9897 or CG32833 either individually or in combination with each other or with CG32834. Consistent with the egg-laying results, the only recurring pattern observed for progeny production was that females with knocked-down levels of CG32834

produced fewer progeny over the course of 24 hours and of 10 days. Thus, the three genes are important for post-mating behaviors.

The case presented in this thesis ultimately uncovers the patterns of gene duplication and co-option: after duplication, the newly arisen gene (the male-specific one) has not only changed the sex-of-expression (from female to male), but has also gained a new function (from a female reproductive protein to an Sfp). This could in fact represent a rapid evolutionary means for creating effective Sfps from female proteins, which in turn could have functionality in the females.

The approach taken in this project was very straightforward and it followed the general rule of solving such questions: i.e. (1) first, one identifies a gene cluster or gene family, (2) one builds a phylogenetic tree of the genes in question, (3) one identifies the origin of gene duplication, (4) one determines the functions of those genes before and after duplication (if molecular and genetic methods are available), (5) one identifies functional shifts after duplication, (6) finally, one pinpoints or proposes the mechanisms driving that shift.

The last step is the most challenging at the moment. Nevertheless, gene duplication is the most common way for new sex-biased genes to arise in the genome, and in particular male-biased genes seem to have arisen frequently through duplication events and have a disproportionate high number of paralogs in the fly and worm genomes. Thus, duplication of previously existing sex-biased or sex-specific genes, along with their regulatory sequences, could represent a way to generate new sex-biased or sex-specific genes, which in turn could gain new functions or exhibit subfunctions relative to the ancestral gene. The case presented in this thesis ultimately uncovers this common mechanism of generating new sex-biased or sex-specific genes, with the added particularity that the male-specific gene arose from the duplication of a female-specific gene. It

is likely that this duplication and diversification reflects a coevolutionary relationship between the sexes.

Coevolution between the sexes is thought to be driven by sexual conflict among male-female interacting proteins, i.e. the different interests and optima between females and males result in an arms race that drives the rapid evolution and coevolution of the interacting reproductive proteins from each sex (40, 60-62). The way sexual conflict plays a role in our case is unclear, but duplication of reproductive proteins followed by co-option of the paralogs in the opposite sex could provide a mechanism to resolve the sexual conflict, if there were one (39, 63-64).

Finally, recent large-scale gene expression studies have shown that gene evolution rates are positively correlated with the tissue specificity of genes, i.e. genes that evolve more rapidly tend to be more tissue specific (65-66). Reproductive genes, especially Sfps, are among the fastest evolving genes discovered to date, so it would make sense that their expression patterns are highly tissue specific, such as the reproductive tract-specific expression pattern of the genes in our case.

The contribution that gene duplication has already provided for evolution is undeniable: without gene duplication, gene subfunctions and new functions would not have arisen, adaptations to changing environments would have been severely limited, etc. Nonetheless, understanding the process of evolution by gene duplication will require detailed molecular characterization of individual gene families, consideration of their physical properties, computational analyses of genomic sequences, use of molecular genetic technology, population genetic modeling, etc. Explorations of the above approaches will help establish evolution by gene duplication as a fundamental principle of evolutionary biology.

CHAPTER V

BIBLIOGRAPHY

1. Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag
2. Bridges, C. B. (1836) *The Bar “gene” a duplication*. Science 83: 210-211
3. Vision, T. J. *et al.* (2000) *The origins of genomic duplications in Arabidopsis*. Science 290: 2114-2117
4. Bowers, J. E. *et al.* (2003) *Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events*. Nature 422: 433-438.
5. Gu, X. *et al.* (2002) *Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution*. Nature Genetics 31: 205-209.
6. McLysaght, A. *et al.* (2002) *Extensive genomic duplication during early chordate evolution*. Nature Genetics 31: 200-204.
7. Postlethwait, J. H. *et al.* (1998) *Vertebrate genome evolution and the zebrafish gene map*. Nature Genetics 18: 345-349.
8. Zhang, J. (2003) *Evolution by gene duplication: an update*. Trends in Ecology and Evolution 18: 292-298.
9. Lynch, M. and Connery, J. S. (2000) *The evolutionary fate and consequences of duplicate genes*. Science 290: 1151-1155.
10. Rodin, S. N. and Riggs, A. D. (2003) *Epigenetic silencing may aid evolution by gene duplication*. Journal of Molecular Evolution 56: 718-729.
11. Lynch, M. *et al.* (2001) *The probability of preservation of a newly arisen gene duplicate*. Genetics 159: 1789-1804.

12. Harrison, P. M. *et al.* (2002) *Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22*. *Genome Research* 12: 272-280.
13. Walsh, J. B. (1995) *How often do duplicating gene evolve new functions?* *Genetics* 139: 421-428
14. Lynch, M. *et al.* (2000) *The probability of duplicate gene preservation by subfunctionalization*. *Genetics* 154: 459-473.
15. Jensen, R. A. (1976) *Enzyme recruitment in the evolution of new function*. *Annual Review of Microbiology* 30: 409-425.
16. Orgel, L. E. (1977) *Gene duplication and the origin of proteins with novel functions*. *Journal of Theoretical Biology* 67:773.
17. Hughes, A. L. (1994) *The evolution of functionally novel proteins after gene duplication*. *Proceedings to the Royal Society of London Series B* 256: 119-124.
18. Force, A. *et al.* (1999) *Preservation of duplicate genes by complementary, degenerative mutations*. *Genetics* 151: 1531-1545.
19. Hughes, A. L. (1999) *Adaptive evolution of genes and genomes*. Oxford University Press.
20. Li, W. H. (1997) *Molecular Evolution*. Sinauer
21. Nei, M. *et al.* (2000) *Purifying selection and birth-and-death evolution in the ubiquitin gene family*. *Proceedings of the National Academy of Sciences of the United States of America* 97: 10866-10871.
22. Piontkivska, H. *et al.* (2002) *Purifying selection and birth-and-death evolution in the histone H4 gene family*. *Molecular Biology of Evolution* 19: 698-697.
23. Ellegren, H. and Parsch, J. (2007) *The evolution of sex-biased genes and sex-biased gene expression*. *Nature Reviews Genetics* 8: 689-698.

24. McLennan, D. A. (2008) *The concept of co-option: why evolution often looks miraculous*. Evolution Education Outreach 1: 247-258.
25. Findlay, G. D. *et al.* (2008) *Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating*. PLoS Biology 6: e178-e188.
26. Kresge, N. *et al.* (2001) *The crystal structure of a fusogenic sperm protein reveals extreme surface properties*. Biochemistry 40: 5407-5413.
27. Vacquier, V. D. *et al.* (1997) *Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence?* Journal of Molecular Evolution 44: 15-22.
28. Clark, N. L. *et al.* (2007) *Duplication and selection on abalone sperm lysin in an allopatric population*. Molecular Biology and Evolution 24: 2081-2090.
29. Parsch, J. *et al.* (2001) *Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of Drosophila simulans*. Genetics 159: 647-657.
30. Nielsen, M. G. *et al.* (2006) *Functional constraint underlies 60 million year stasis of dipteran testis-specific β -tubulin*. Evolution & Development 8: 23-29.
31. Raff, E. C. *et al.* (2000) *Conserved axoneme symmetry altered by a component β -tubulin*. Current Biology 10: 1391-1394.
32. Nielsen, M. G. *et al.* (2010) *Tubulin evolution in insects: gene duplication and subfunctionalization provide specialized isoforms in a functionally constraint gene family*. BMC Evolutionary Biology 10: 113-123.
33. Baker, R. H. *et al.* (2012) *Gene duplication, tissue-specific gene expression and sexual conflict in stalk-eyed flies (Diopsidae)*. Philosophical Transactions of the Royal Society B 367: 2357-2375.

34. Gao, G. *et al.* (2011) *Paternal imprint essential for the inheritance of telomere identity in Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 108: 4932-4937.
35. Gallach, M. *et al.* (2010) *Analyses of nuclearly-encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexual antagonistic conflict in Drosophila*. *Genome Biology and Evolution* 2: 835-850.
36. Danshina, P. V. *et al.* *Phosphoglycerate kinase 2 (PGK2) is essential for sperm function and male fertility in mice*. *Biology of Reproduction* 82: 136-145.
37. Belote, J. M. *et al.* (2009) *Duplicated proteasome subunit genes in Drosophila and their roles in spermatogenesis*. *Heredity* 103: 23-31.
38. Tracy, C. *et al.* (2010) *Convergently recruited nuclear transport retrogenes are male biased in expression and evolving under positive selection in Drosophila*. *Genetics* 184: 1067-1076.
39. Gallach, M. and Betran, E. (2011) *Intralocus sexual conflict resolved through gene duplication*. *Trends in Ecology and Evolution* 26: 222-228.
40. Haerty, W. *et al.* (2007) *Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila*. *Genetics* 177: 1321-1335.
41. Manchini, E. *et al.* (2011) *Molecular evolution of a gene cluster of serine proteases expressed in the Anopheles gambiae female reproductive tract*. *BMC Evolutionary Biology* 11: 72-89.
42. Manchini, E. *et al.* (2011) *Molecular characterization and evolution of a gene family encoding male-specific reproductive proteins in the African malaria vector Anopheles gambiae*. *BMC Evolutionary Biology* 11: 292-308.

43. Tian, X. *et al.* (2009) *Evolution and functional divergence of NLRP genes in mammalian reproductive systems*. BMC Evolutionary Biology 9: 202-214.
44. Kleineidam, R. G. *et al.* (1999) *Seminal-type ribonuclease genes in ruminants, sequence conservation without protein expression?* Gene 231: 147-153.
45. Trabesinger-Ruef, N. *et al.* (1996) *Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function?* FEBS Letters 382: 319-322.
46. Clauss, A. *et al.* (2005) *The evolution of a genetic locus encoding small serine proteinase inhibitors*. Biochemistry and Biophysical Research Communications 333: 383-389.
47. Karn, R. C. *et al.* (2008) *Adaptive evolution in rodent seminal vesicle secretion proteins*. Molecular Biology and Evolution 25: 2301-2310.
48. Kelleher, E. S. *et al.* (2007) *Gene duplication and adaptive evolution of digestive proteases in Drosophila arizonae female reproductive tracts*. PLoS Genetics 3: e148-e155.
49. Kelleher, E. S. and Pennington, J. E. (2009) *Protease gene duplication and proteolytic activity in Drosophila female reproductive tracts*. Molecular Biology and Evolution 26: 2125-2134.
50. Kelleher, E. S. and Markow, T. A. (2009) *Duplication, selection and gene conversion in a Drosophila mojaveensis female reproductive protein family*. Genetics 181: 1451-1465.
51. Wagstaff, B. J. and Begun, D. J. (2007) *Adaptive evolution of recently duplicated accessory gland protein genes in desert Drosophila*. Genetics 177: 1023-1030.
52. Sartain, C. V. *et al.* (2011) *The poly(A) polymerase GLD2 is required for spermatogenesis in Drosophila melanogaster*. Development 138: 1619-1629.
53. Cui, J. *et al.* (2008) *Wispy, the Drosophila homolog of GLD-2, is required during oogenesis and egg activation*. Genetics 178: 2017-2029.

54. Benoit, P. *et al.* (2008) *PAP- and GLD-2-type poly(A) polymerases are required sequentially in cytoplasmic polyadenylation and oogenesis in Drosophila*. *Development* 135: 1969-1979.
55. Arbeitman, M. N. *et al.* (2004) *A genomic analysis of Drosophila somatic sexual differentiation and its regulation*. *Development* 131: 2007-2021.
56. Chintapalli, V. R. *et al.* (2007) *Using FlyAtlas to identify better Drosophila melanogaster models of human disease*. *Nature Genetics* 39: 715-720.
57. Prokupek, A. M. *et al.* (2009) *Transcriptional profiling of the sperm storage organs in Drosophila melanogaster*. *Insect Molecular Biology* 18: 465-475.
58. Allen, A. K. and Spradling, A. C. (2008) *The Sfl-related nuclear hormone receptor Hr39 regulates Drosophila female reproductive tract development and function*. *Development* 135: 311-321.
59. Ravi Ram, K. and Wolfner, M. F. (2007) *Seminal influences: Drosophila Acps and the molecular interplay between males and females during reproduction*. *Integrative and Comparative Biology* 47: 427-445.
60. Wolfner, M. F. (2002) *The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in Drosophila*. *Heredity* 88: 85-93.
61. Wolfner, M. F. (2009) *Battle and ballet: molecular interactions between the sexes in Drosophila*. *Heredity* 100: 399-410.
62. Sirot, L. K. *et al.* (2009) *Molecular social interactions: Drosophila melanogaster seminal fluid proteins as a case study*. *Advances in Genetics* 68: 23-56.
63. Gallach, M. *et al.* (2011) *Gene duplication and the genome distribution of sex-biased genes*. *International Journal of Evolutionary Biology* ID989438

64. Connallon, T. and Clark, A. G. (2011) *The resolution of sexual antagonism by gene duplication*. Genetics 187: 919-937.
65. Winter, E. E. *et al.* (2004) *Elevated rates of protein secretion, evolution, and disease among tissue-specific genes*. Genome Research 14: 54-61.
66. Zhang, L. *et al.* (2004). *Mammalian housekeeping genes evolve more slowly than tissue-specific genes*. Molecular Biology and Evolution 21: 236-239.